PATENT ABSTRACTS OF JAPAN

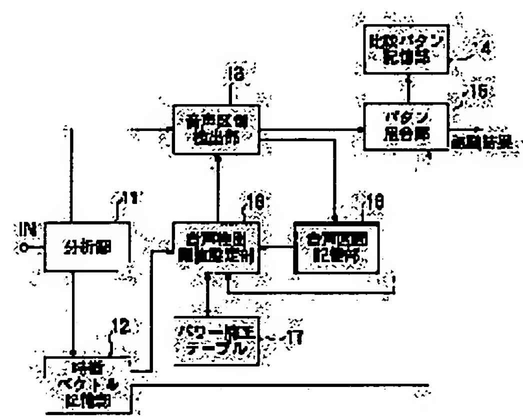
(11)Publication number : 07-092989
(43)Date of publication of application : 07.04.1995

(51)Int.CI. G10L 3/00

(21)Application number : 05-236880
(22)Date of filing : 22.09.1993
(71)Applicant : OKI ELECTRIC IND CO LTD
(72)Inventor : MIKI TAKASHI

(54) SPEECH RECOGNIZING METHOD

(57)Abstract:
PURPOSE: To secure the detection of a speech input period in speech recognition.
CONSTITUTION: An analytic part 11 calculates a 1st feature vector and power representing the feature of a signal and stores them in a feature vector storage part 12. A speech section detection part 13 detects the speech input period from the power by using a threshold value for speech detection. A pattern matching part performs the speech recognition by comparing the 1st feature vector of the input signal inputted in the speech input period with a 2nd feature vector which is stored in a comparison pattern and represents the feature of the speech to be recognized. A speech detection threshold value setting part 18 complements noise and vocalization environment on the basis of the result of the speech recognition by using a coefficient stored in a power correction table 17 to update the threshold value for speech detection.



LEGAL STATUS

[Date of request for examination]	17.02.1997
[Date of sending the examiner's decision of rejection]	18.01.2000
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	
[Date of registration]	
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	
[Date of extinction of right]	

Copyright (C); 1998,2000 Japan Patent Office

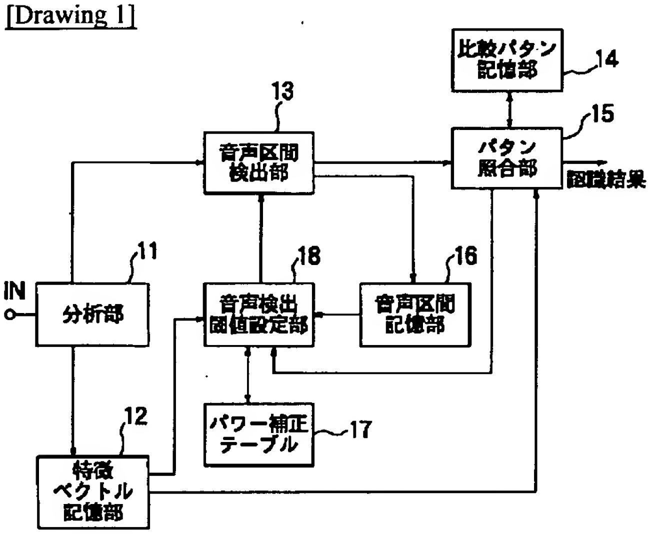
* NOTICES *

Japan Patent Office is not responsible for any damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

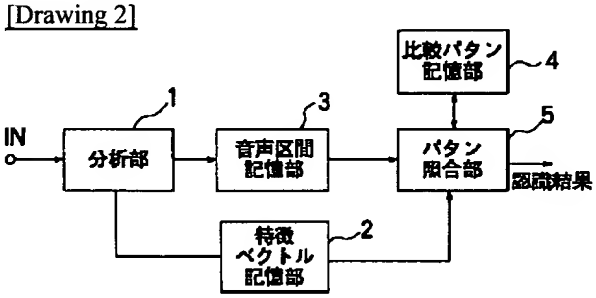
DRAWINGS

[Drawing 1]



本発明の音声認識方法を実施する音声認識装置

[Drawing 2]



従来例の音声認識装置

[Drawing 3]

母音	あ	い	う	え	お
パワー比	1.00	0.64	0.72	0.90	0.75

一般的な母音間のパワー比

[Translation done.]

* NOTICES *

Japan Patent Office is not responsible for any damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] It is the configuration block view showing the example of equipment which enforces the speech recognition method of this invention.
[Drawing 2] It is the configuration block view showing the voice recognition unit of the conventional example.
[Drawing 3] It is drawing showing the power ratio between general vowels.
[Description of Notations]
1 11 Analyzor
2 12 Feature-vector storage section
3 13 Voice section detecting element
4 14 Comparison pattern storage section
5 15 Pattern-matching section
16 Voice Section Storage Section
17 Power Amendment Table
18 Voice Detection Threshold Setting Section

[Translation done.]

* NOTICES *

Japan Patent Office is not responsible for any damages caused by the use of this translation.

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[Industrial Application] this invention relates to the voice section method of detection in the voice recognition unit used as a computer, a control unit, and an input means of various devices in addition to this.
[0002]
[Description of the Prior Art] Conventionally, as technology of such a field, there were some which are indicated by the following reference, for example.
Reference 1; JP,62-73298,A drawing 2 is the configuration block view showing the voice recognition unit of the conventional example. The input terminal IN which inputs the input signal S which the conventional recognition equipment of the voice recognition unit of drawing 2 as shown in reference 1 was typical, and contained the sound signal, The analysis section 1 which analyzes the power Pt of the 1st feature vector which expresses the feature of an input signal S for every fixed period, and the input signal S of each frame t of every which calls the input signal S a frame, The feature-vector storage section 2 which stores the analysis result of these 1st feature vectors, The voice section detecting element 3 which asks for the voice input period when it detects in whether a sound signal is in an input signal from each power Pt of an input signal, and whether there is nothing, and voice is inputted, The comparison pattern storage section 4 in which the 2nd time series signal which consisted of the 2nd feature vector which is a feature vector of the voice for recognition beforehand is stored, It has the pattern-matching section 5 which compares the time series signal of the 2nd feature vector stored in the comparison pattern storage section 4 with the 1st time series signal which consisted of the 1st feature vector stored in the feature-vector storage section 2.
[0003] Next, operation of the recognition equipment of drawing 2 is explained. The analyzor 1 calculates the 1st feature vector with the power Pt of this input on every [of the short section] frame t to the input signal S inputted from the input terminal IN. Here, as the 1st feature vector, the spectrum parameter expressing frequency spectrum is common. The 1st feature vector called for by the analyzor 1 is saved one by one in the feature-vector storage section 2. The voice section detecting element 3 determines a voice input period from each power Pt of the input signal S in each frame t. That is, the voice section detecting element 3 finds out the section exceeding the threshold Psh for voice detection more than the predetermined period with Power Pt, and memorizes the section as the voice candidate section. Furthermore, the voice section detecting element 3 finds out the section where the signal S with which Power Pt exceeds the threshold Psh for voice detection more than predetermined time from the voice candidate section is not inputted, and considers the time as an audio end. The voice section detecting element 3 determines the start edge and termination of voice input from the voice candidate section memorized after the end of voice input. Let the start edge of voice input be the earliest frame on a time target that is among the voice candidate section. Moreover, let termination of voice input be the latest frame on a time target that is among the voice candidate section. The pattern-matching section 5 calculates both degree of similar after the determination of a voice input period by collating the 1st time series signal (following input voice pattern) which consisted of the 1st feature vector from the start edge of voice input to termination, and two or more 2nd time series signals (henceforth a comparison pattern) which consisted of the 2nd average feature vector of the voice for recognition stored in the comparison pattern storage section 4. The word name given to the comparison pattern which gives the maximum degree of similar, and which expresses a word, for example is outputted as a recognition result.
[0004] In addition, the important threshold Psh for voice detection has the method of deciding in (i) - (iii) a procedure below by voice section detection processing.
(i) Are in the state where level measurement voice of a background noise is not inputted, namely, it is the average noise power Pnoise of the average of the power Pt of only noise. It is measured.
(ii) Maximum Pvoice of the power Pt in voice level measurement voice input, i.e., the maximum voice power, It measures.
(iii) Psh is decided for the threshold for threshold setting voice detection by (1) formula.
Psh=0.03xPvoice+0.97xPnoise ...... (1)
[0005]
[Problem(s) to be Solved by the Invention] However, the following technical problems occurred in the conventional speech recognition method. in order to set up the threshold Psh for voice detection, only noise is inputted beforehand -- **** -- two or more frames to average noise power Pnoise The maximum voice power Pvoice in the section which consisted of frames into which the sound signal is inputted It is necessary to measure. However, noise level changes every moment and audio

power also changes. Then, average noise power Pnoise And the maximum voice power Pvoice It needed to measure again suitably and the threshold Psh for voice detection needed to be updated. However, it is the maximum voice power Pvoice to a user at high frequency. It is the phonation at the time of recognition operation to the maximum voice power Pvoice preferably to force it the phonation for measurement. The processing which measures again is mainly used. On the other hand, it is the maximum voice power Pvoice. There is a close relation to the phoneme contained in the voice section. for example, the vowel "**" power Pt -- most -- strong -- vowel" -- it is and the power Pt of "is the smallest Consequently, the maximum power Pvoice of the section where voice is inputted It differs for every word. However, since these differences cannot be taken into consideration by the conventional method of using the phonation at the time of recognition, it is the maximum voice power Pvoice. The value lacked in reliability. Consequently, the threshold for voice detection may become unsuitable and had become the cause of a fall of voice detection precision. As a technical problem which the aforementioned conventional technology had, this invention offers the speech recognition method by which reliability of voice input period detection was solved about the low point.

[0006]

[Means for Solving the Problem] Feature-vector calculation processing in which the 1st feature vector which expresses the power of this input signal and the feature of the input signal for every frame to an input signal is computed and saved in order that this invention may solve the aforementioned technical problem, Voice section detection processing in which detect the aforementioned frame which compares the threshold for voice detection with the aforementioned power for every frame, and contains the sound signal in the aforementioned input signal, and the voice input period when voice is inputted is detected, The 1st time series signal which consisted of two or more 1st feature vectors of the above of the aforementioned voice input period, The 2nd time series signal which consisted of two or more 2nd feature vectors which corresponded to two or more voice for recognition, respectively is compared. The following processings are performed in the speech recognition method of performing speech recognition processing which searches for the voice name for recognition corresponding to the time series signal which was most similar to the time series signal of the above 1st among system multiple-message-transmission numbers at the 2nd time of this. Namely, the maximum voice power calculation processing in which it asks for the maximum voice power of the aforementioned power in the latest past voice input period among the voice input periods which speech recognition processing already ended, It asks for the correction factor corresponding to this voice for recognition from the voice name for recognition called for from the latest voice input period of the aforementioned past. the aforementioned maximum voice power by this correction factor The amendment maximum voice power amendment processing, Background-noise power presumption processing in which the size of noise is presumed from the aforementioned power of periods other than the aforementioned voice input period, The voice detection threshold update process which updates the aforementioned threshold for voice detection used for the aforementioned future voice section detection processings is performed from the aforementioned maximum voice power amendment processing and a background-noise power presumption processing result.

[0007]

[Function] According to this invention, since the speech recognition method was constituted as mentioned above, the 1st feature vector which expresses the power of an input signal and the feature of the input signal for every frame is computed and saved by feature-vector.calculation processing. By voice section detection processing, the threshold for voice detection is compared with the aforementioned power for every frame, the frame which contains the sound signal in the input signal is detected, and a voice input period is called for. If a voice input period is called for, the voice for recognition corresponding to the sound signal will be called for by speech recognition processing. Then, the maximum voice power of the power in the latest past voice input period is called for by the maximum voice power calculation processing among the voice input periods which speech recognition processing already ended. The maximum voice power amendment processing asks for the correction factor corresponding to this voice for recognition from the voice name for recognition called for from the voice input period, and amends the maximum voice power. The size of the noise of periods other than a voice input period is presumed, and the threshold for voice detection used for the next recognition processing by voice detection threshold update process is updated by background-noise power presumption processing from these maximum voice power amendment processing and a background-noise power presumption processing result. Therefore, the aforementioned technical problem is solvable.

[0008]

[Example] Drawing 1 is the configuration block view showing the example of equipment which enforces the speech recognition method of this invention. The input terminal IN which inputs the input signal S which the equipment of drawing 1 is a voice recognition unit which recognizes the inputted word, and contains a sound signal, The 1st feature vector which expresses the feature of an input signal S for the input signal S for every frame Pt, respectively, and the analyzor 11 which computes the power Pt of the input signal S for every frame Pt, respectively, The feature-vector storage section 12 which saves these 1st feature vectors and the calculation result of Power Pt, The voice section detecting element 13 which asks for the voice input period when it detects in whether a sound signal is in an input signal S from the power Pt, and whether there is nothing, and voice is inputted, The comparison pattern storage section 14 in which the 2nd time series signal which consisted of the 2nd feature vector which is a feature vector of the voice for recognition (word) beforehand is stored, It has the pattern-matching section 15 which compares the time series signal of the feature vector stored in the comparison pattern storage section 14 with the 1st time series signal which consisted of the 1st feature vector. Furthermore, this equipment has formed the voice detection threshold setting section 18 which computes the voice detection threshold for updating from the

information on the voice section storage section 16 which stores the information on the voice input period from the voice section detecting element 13, the power amendment table 17 on which the data for the maximum voice power amendment for every voice for recognition were stored, and Power Pt and a voice input period, and is supplied to the voice section detecting element 13.

[0009] Next, operation of the voice recognition unit of drawing 1 is explained. To the input signal S inputted from the input terminal IN, the analyzor 11 computes Power Pt and the 1st feature vector of this input on every [of the short section] frame t, and generates the time series signal of the 1st feature vector. In the calculation method of the 1st feature vector, the method using two or more band pass filter groups from which center frequency differs little by little, the method using the analysis of a spectrum by FFT (fast Fourier transform), etc. can be considered. Here, an example is explained for how to use a band pass filter group. In the analyzor 11, an input signal S is changed into an analog signal or a digital signal, and each band pass filter in the analyzor 11 extracts two or more frequency components of an input signal S. Thus, the sequence of the data which were able to be distributed by each band pass filter is called a channel. The output signal of the filter for every channel is rectified, and every frame t is asked for the average for every output signal of a filter. This calculated average is called band power and the band power of the j-th channel is expressed in the t-th frame as Ftj. Next, the analyzor 11 computes Power Pt for every frame. Calculation of Power Pt is computed by the following (2) formulas. Furthermore, two or more 1st feature vectors Gtj are computed by (4) formulas from each band power Ftj.

[0010]

[Equation 1]

$$Pt = \log\left(\sum_{j=1}^{p} Ftj\right) \quad \cdots\cdots\cdots (2)$$

$$Gt = (Gt1,\ Gt2,\ \cdots,\ Gtp\text{-}1,\ Gtp) \quad \cdots\cdots\cdots (3)$$

但し、p ：チャネル数

$$Gtj = \log(Ftj) \quad \cdots\cdots\cdots (4)$$

The 1st feature vector and Power Pt are saved in the feature-vector storage section 12 by N frames. That is, feature-vector calculation processing is performed in the analyzor 11 and the feature-vector storage section 12. The voice section detecting element 13 performs voice section detection processing in which the period when the sound signal is continuously inputted by the set-up threshold Psh for voice detection is detected, based on Power Pt. That is, the frame A of the start edge of voice input and the termination frame B are determined. The voice section detecting element 3 finds out the section exceeding the threshold Psh for voice detection more than the predetermined period with Power Pt, and memorizes the section as the voice candidate section. Furthermore, the voice section detecting element 3 finds out the section where the signal S with which Power Pt exceeds the threshold Psh for voice detection more than predetermined time is not inputted after voice candidate section detection, and considers the time as an audio end. From the memorized voice candidate section, the voice section detecting element 3 determines the start edge A and Termination B of voice input. Let the start edge of voice input be the earliest frame on a time target that is among the voice candidate section. Moreover, let termination of voice input be the latest frame on a time target that is among the voice candidate section. The information on these voice input periods is stored in the voice section storage section 16. If the voice section is determined, the pattern-matching section 15 will calculate both degree of similar by collating two or more comparison patterns, the 1st time series signal, i.e., the input voice pattern, which consisted of the 1st feature vector from the start edge of voice input to termination, of the 2nd time series signal which consisted of the 2nd average feature vector of the voice for recognition stored in the comparison pattern storage section 4. The word name r given to the comparison pattern of the word which gives the maximum degree of similar is outputted as a recognition result.

[0011] The voice detection threshold setting section 18 carries out an updating setup of the subsequent thresholds for voice detection at the following step 1 - Step 4 after the above speech recognition operation end.

(Step 1) The maximum voice power Pvoice Two or more power Pt to the maximum voice power Pvoice which corresponds between frame AM-BM of the latest voice input section which speech recognition ended among the power Pt memorized by the calculation processing feature-vector storage section 12 (Mth voice section) It asks by (5) formulas. In addition, the information on a voice input period is stored in the voice section storage section 16.

Pvoice =max {Pt} ...... (5) AM<=t<=BM (Step 2) The word name r called for in the maximum voice power amendment processing pattern-matching section 15 is used, and it is the maximum voice power Pvoice. Amendment. That is, the coefficient Tr corresponding to each word name r is beforehand stored in the power amendment table 17, respectively, and it is the maximum voice power Pvoice by the coefficient Tr. It is rectified by (6) formulas and the amendment maximum voice power Pvoice1 is called for.

Pvoice1 = Pvoice+Tr [r] ...... (6) (Step 3) Background-noise power Pnoise The average of Power Pt is calculated from the sections other than the sound signal input section determined by the presumed processing voice section detecting element 13, and it is the background-noise power Pnoise by this. It is presumed. The concrete calculation range is restricted by length N of the storage region of Power Pt. If power memorized by the feature-vector storage section 12 is set to Pt and the entire interval

memorized by the voice section storage section 16 is made into frame AM-BM (Mth voice section) from frames A1-B1 (1st voice section), it is the background-noise power Pnoise. It is computed by (7) and (8) formulas.
[0012]
[Equation 2]

$$N_o = N - \sum_{m=1}^{M} \sum_{t=Am}^{Bm} \quad \cdots\cdots\cdots (7)$$

$$Pnoise = \left(\sum_{n=1}^{N} pt - \sum_{m=1}^{M} \sum_{t=Am}^{Bm} pt \right) / N_o \quad \cdots\cdots\cdots (8)$$

(Step 4) The update process amendment maximum voice power Pvoice1 of the voice detection threshold Psh, and background-noise power Pnoise from -- the voice detection threshold Psh is updated by (9) formulas Psh=0.03xPvoice1+0.97xPnoise ...... (9) This updated threshold Psh for voice detection is used at the time of next speech recognition. Furthermore, there is also the method of updating gradually by the forgotten type learning method which does not set up the threshold Psh for voice detection by one phonation, but is shown by (10) formulas.
Psh=aPsh(n-1) +(1-a) Psh* ...... (10) However, Psh(n);n time threshold Pshfor voice detection * ; Threshold a for voice detection calculated by (9) formulas from the n-1st phonation ; It is necessary to set the power correction factor Tr to processing in the voice detection threshold setting section 18 more than an updating coefficient beforehand. A setup of a correction factor Tr is performed at following Step A and following Step B.
[0013] Step A It asks for the standard maximum voice power of the word for voice power setting recognition by a certain method. For example, there is the method of calculating the maximum voice power by uttering the word for recognition several times. Calculation of the maximum voice power is calculated in the same procedure as the case of the voice input at the time of previous speech recognition. The sound signal of the word name r inputted from the voice input terminal IN is changed into the time series signal and Power Pt (r) of a feature vector by the analyzor 11. In a voice section detecting element, the start edge frame Ar and the termination frame Br of the voice section, i.e., voice input, are determined based on Power Pt (r), and it is the maximum voice power Pvoice of Power Pt (r). (r) is (11) formulas and it is ********.
Pvoice (r) =max {Pt (r)} ...... (11) Ar<=t<=Br The maximum of voice power is calculated from further two or more phonation, and it is Pvoice about the average. It is good also as (r). Moreover, there is also a method of presuming standard voice power using the knowledge of phonetics from the pronunciation notation of the word for recognition etc. as an example of the method of asking for the standard maximum voice power. The maximum phonation power of a certain word is almost equal to the maximum power of a vowel portion. It does not depend on an individual mostly, but the power ratio between different-species vowels is good noting that it is fixed. Therefore, if the pronunciation sequence of the word for recognition is known, it can ask for the maximum power ratio between words by calculation. Drawing 3 is drawing showing the power ratio between general vowels. for example," -- yet -- "" -- ******"" -- the maximum voice power ratio of each word [are and] flower-stalk" -- vowel" -- **"" -- obtaining -- "" -- it becomes the power ratio 1:0.72:0.75 of " It is Pvoice as it is about this power ratio. (r) Then, it is good.
[0014] Step B The voice power setting maximum voice power Pvoice The power correction factor Tr (r) is set up for every word for recognition from (r). The power correction factor Tr (r) is the maximum Pvoice of all the words for recognition. It is the difference of the average and each correction factor Tr (r), and asks by (12) and (13) formulas.
[Equation 3]

$$Pav = \sum_{r=1}^{R} Pvoice(r) / R \quad \cdots\cdots\cdots (12)$$

R：認識対象語の語数

$$Tr(r) = Pav - Pvoice(r) \quad \cdots\cdots\cdots (13)$$

Ｔｒ(r) の値はパワー補正テーブル１７に記憶される。

As mentioned above, in this example, it can respond to the power of the noise environment and voice which always change, and the threshold for voice detection can always be updated. Moreover, the power characteristics which, originally [of the word for recognition itself], have the updated threshold for voice detection are taken into consideration. Therefore, the error of detection of a voice input period is reduced and reliable detection is carried out. As a result, a high speech recognition performance can be obtained. In addition, this invention is not limited to the above-mentioned example, but various deformation is possible for it. As the modification, there is the following, for example.

(1) Although the voice for recognition is considering as the word, it is not limited to a word but do so the effect as an example that this invention is the same also as language by means of which phonation of a speaker, one pronunciation, or the word stood in a row, according to the purpose.
(2) Although it asks for the voice candidate section and is asking for the start edge and termination of voice input, to the algorithm which asks for the start edge and termination of these voice input, various deformation is possible.
(3) The composition of a voice recognition unit is not limited to drawing 1 , for example, the power amendment table 17 or the comparison pattern storage section 14 is good also as composition installed in external another equipment and an external storage.
[0015]
[Effect of the Invention] As explained to the detail above, according to this invention, based on the speech recognition result in the latest past voice input period, processing which updates the threshold for voice detection is carried out. That is, the power characteristics which, originally [of the word for recognition itself], have the updated threshold for voice detection are taken into consideration. Therefore, while corresponding to the power of the noise environment and voice which always change, the error of detection of a voice input period is reduced and reliable detection is carried out. As a result, a high speech recognition performance can be obtained.

[Translation done.]

* NOTICES *

Japan Patent Office is not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer.So the translation may not reflect the original precisely.
2. **** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] Feature-vector calculation processing in which the 1st feature vector which expresses the power of this input signal and the feature of the input signal for every frame to an input signal is computed and saved, Voice section detection processing in which detect the aforementioned frame which compares the threshold for voice detection with the aforementioned power for every frame, and contains the sound signal in the aforementioned input signal, and the voice input period when voice is inputted is detected, The 1st time series signal which consisted of two or more 1st feature vectors of the above of the aforementioned voice input period, The 2nd time series signal which consisted of two or more 2nd feature vectors which corresponded to two or more voice for recognition, respectively is compared. In the speech recognition method of performing speech recognition processing which searches for the voice name for recognition corresponding to the time series signal which was most similar to the time series signal of the above 1st among system multiple-message-transmission numbers at the 2nd time of this The maximum voice power calculation processing in which it asks for the maximum voice power of the aforementioned power in the latest past voice input period among the voice input periods which speech recognition processing already ended, It asks for the correction factor corresponding to this voice for recognition from the voice name for recognition called for from the latest voice input period of the aforementioned past. the aforementioned maximum voice power by this correction factor The amendment maximum voice power amendment processing, Background-noise power presumption processing in which the size of noise is presumed from the aforementioned power of periods other than the aforementioned voice input period, The speech recognition method characterized by performing the voice detection threshold update process which updates the aforementioned threshold for voice detection used for the aforementioned future voice section detection processings from the aforementioned maximum voice power amendment processing and a background-noise power presumption processing result.

[Translation done.]